Linguistic and philosophical considerations on Bayesian semantics

Daniel Lassiter
Stanford University

Abstract Recently many theories of the semantics and pragmatics of deontic modals have relied on analogies with decision theory. One particularly straightforward way to make this connection is to build the semantics around degree scales, where the degrees in question involve expected (moral, personal, etc.) value. This "Bayesian" semantics has a number of attractive features involving information-sensitivity, grammatical gradability, and solutions to logical puzzles which plague standard theories of deontic semantics. However, its viability has been questioned by several authors who argue that it is insufficiently expressive empirically or that it builds philosophically suspect assumptions into the semantic theory. These authors have advocated moving instead to a much weaker semantic theory which can express the predictions of the Bayesian theory among many others. In this paper I survey some of the arguments in favor of the Bayesian semantics and respond to objections, arguing that the empirical problems can be met, that the philosophical objections are unconvincing, and that, absent compelling empirical arguments, empirical data and methodological considerations favor a strongly predictive theory of deontic semantics over one with much greater expressivity.

Keywords: deontic modals, decision theory, scalar reasoning, attitude contexts, philosophy of linguistics

For the past several years I have found myself advocating what might be called a "Bayesian" semantics for the deontic interpretations of modal expressions such as *ought, should, good, bad, may, must*, and their modified forms (*better*, etc.). On the version of this account that I have in mind, deontic modals are not quantifiers over possible worlds; instead, they have a lexical semantics structured around degree scales, just as gradable adjectives such as *heavy/light, tall/short*, and *full/empty* are usually thought to (Kennedy & McNally 2005, etc.). Deontic scales are formally identical to the expected utility scales used in Bayesian decision theory, but the function assigning values to states is not necessarily interpreted as representing personal utility — it could be, as in the case of teleological modals, but it could also represent other kinds of value, such as moral value. I have argued that this approach accounts better than its main rivals for a number of empirical phenomena including information-sensitivity, non-monotonicity, and grammatical gradability (Lassiter 2011, to appear, 2014c).[1]

1 The Bayesian semantics made its first appearance (to my knowledge) in a brief paper by Jeffrey (1965a), and was fleshed out and empirically motivated by Jackson (1991), formalized by Goble (1996), and integrated into a compositional semantics for English, with special attention to the theory of gradability, by Lassiter (2011, 2014c, to appear). Related ideas can be found in Cariani 2009; Wedgwood ms.

Note that the label "Bayesian" is not much used in these works, since the connection with Bayesian decision theory is mostly formal in nature. But the term is used in subsequent reactions (e.g. Yalcin 2012a; Cariani 2015), and I find that

In this paper I will summarize several of the arguments for the Bayesian semantics and discuss challenges that have been posed in four recent papers, two of which are in this volume: Carr 2012; Charlow 2015, and Cariani 2014, 2015. The challenges involve both empirical and high-level philosophical objections, and they have been taken to motivate weakening the semantics so that it can mimic the predictions of the Bayesian theory, or of other existing semantic theories, as desired. For obvious reasons, it is not possible to show on empirical grounds that a strictly weaker theory is incorrect. However, I will argue that the objections that Carr, Charlow, and Cariani put forward are not compelling. For the most part, the problems discussed are not even specific problems for the Bayesian semantics, but rather very general issues that arise for anyone who is trying to do lexical semantics within the framework of compositional model-theoretic semantics, and we can reasonably expect that solutions to these general problems will explain the special cases which involve deontic modals. Given this, general methodological considerations favor adopting a more restrictive semantics over a more expressive one.

An important exception to this characterization is Charlow's (2015) objection involving the deontic analogue of an evidential vs. causal decision theory: e.g., if a behavior and a disease have a common genetic cause, *should* individuals refrain from the behavior? Charlow argues that, to the extent that there is variation among English speakers in their judgments on this point, no single decision theory can be implicated in the semantics of *ought*. However, this argument assumes that the distinction between causal and evidential decision theory must be located in the rule used to compute expected values. As I discuss in section 3, there is a prominent version of decision theory according to which the causal/evidential divide is located in the structure of a probabilistic model and the way that actions/interventions are represented. On this construal, the Bayesian semantics can generate either kind of predictions, depending on features of the probability distribution that is given as a semantic parameter. If this general construal of causal and evidential decision theory is successful, then, Charlow's argument does not motivate adopting a more expressive semantics.

A final category of objections involve real and apparent limitations in the Bayesian theory's ability to model certain conceivable kinds of reasoning about obligation. The real limitations involve MaxiMax or MaxiMin choice rules, but I will argue that the mere *conceivability* of someone employing such a choice rule in deontic reasoning is not a sufficient reason to weaken the semantic theory dramatically. For a compelling objection, empirical evidence that actual people do so would be needed, and this evidence is conspicuously absent. The apparent limitation involves, for example, non-consequentialist judgments in trolley problems. While there is abundant evidence that many people have such judgments, the Bayesian theory can handle it straightforwardly by locating variation in the value parameter: non-consequentialist intuitions are associated with value functions which attach positive or negative value to the *fact that a particular individual takes a particular kind of action*. In other words, the Bayesian theory is *not* a semanticization of consequentialist ethics. It is a logical and grammatical theory, and it stands or falls with the descriptive adequacy of its validities, and with its ability to mesh with good theories of natural language syntax and pragmatics.

it is not inappropriate as long as it is clear that this semantic theory does not stand or fall with the empirical adequacy of a Bayesian theory of human decision-making. At most, there is a philosophical connection in Jackson's (1991: 464) observation that a moral value function can be thought of as describing how people ought to be motivated in their subjective decision-making.

It may be useful for me to make clear from the beginning what my goals and presuppositions are. I am a linguist and a cognitive scientist, and I am primarily interested in understanding language, concepts, and reasoning as aspects of human cognition. Metaphysical questions are relevant to these research interests, but only to the extent that humans have metaphysical beliefs and assumptions which influence their cognition and behavior. The question of the ultimate truth of one or another set of metaphysical claims simply does not arise in this enterprise. Probably in part because of this orientation, I am unsure to what extent I disagree philosophically with critics of the Bayesian semantics. (I don't have any particular objection to Charlow's expressivism, for example, but I also don't know whether the distinction between truth-conditional and expressivist frameworks does any work in a cognitively-oriented theory of meaning and communication.) In any case, I hope that it will at least emerge that the criticisms considered here are not special problems for the Bayesian semantics, and that this theory does a pretty good job of regimenting the grammatical and inferential features of deontic modals. Indeed, I think that it does this job much better than any other theory that has been spelled out with a comparable level of specificity and predictive power.

1 Empirical and theoretical motivation for Bayesian semantics

As I am using the term, the Bayesian semantics relies on two basic claims: that deontic modals have a semantics built around **scales**, and that the scales in question have a particular logical structure. The motivation for building the semantics around degree scales rather than (just) quantification over possible worlds goes back to observations made by Lewis (1973) and Kratzer (1981, 1991): many modal expressions have graded meanings and combine grammatically with degree modifiers and form comparatives and equatives. This is true of epistemic, deontic, teleological, and bouletic modals alike, but we will focus here on deontic expressions. Compare (1) to (2) and (3) to (4), illustrating parallel grammatical gradability among non-modal adjectives and verbs.

(1) Murder is worse/better than jaywalking.

(2) Bill is sadder/happier than Mary.

(3) You ought very much to leave.

(4) Bill likes very much to play golf.

Lassiter (2011: §5) gives much naturalistic evidence of grammatical gradability of deontic modals and related expressions, and additional evidence (involving, for example, neg-raising behavior and focus sensitivity) which supports the claim that many of these expressions have a scalar semantics. There are some notable exceptions: while *ought, should, good, bad, required*, and *supposed to* are gradable, there is no clear evidence for the gradability of *may*. It is still unclear what the situation is for *must*, or *have to*: it's usually thought that they are not gradable, but Portner & Rubinstein (this volume) argue that they are, and that their limited gradability is closely related to the limitations on gradability of extreme adjectives like *huge* and *gorgeous* (Morzycki 2012). Regardless of how this empirical question comes out, gradability and scalarity are not the same thing: gradability is a grammatical property of expressions, while scalarity is a property of the model-theoretic objects that expressions denote (see Lassiter 2014b for discussion). It might turn out that deontic modals that are not gradable nevertheless have a scalar component of their meanings, but the thresholds that they invoke cannot be bound by other operators.

If this is correct, the next question to ask is what kinds of scales are relevant. In the recent literature on degree expressions a great deal of attention has been devoted to **scale structure**, i.e., the formal properties of degree scales (see surveys in Lassiter 2015; Morzycki to appear). The bulk of attention in this literature has been devoted to two issues. First, there is the question of how to model the difference between one-dimensional adjectives like *tall/short* and higher-dimensional adjectives like *healthy/sick*, where many factors go into the determination of the degree of the property that an individual possesses (Kamp 1975; Bierwisch 1989; Sassoon 2013). Second, there is the question of whether a scale has or lacks a lower and upper bound, which has been argued to influence the vagueness of adjectives and their potential to combine with degree modifiers such as *slightly* and *completely* (Hay, Kennedy & Levin 1999; Rotstein & Winter 2004; Kennedy & McNally 2005; Kennedy 2007).

In Lassiter 2011, 2014c, to appear I build on the insights of Measurement Theory (Krantz, Luce, Suppes & Tversky 1971) to identify a third parameter of scalar variation: the **logical part-whole structure** of scales, i.e., systematic relationships between the degree of a property that some object possesses and the degree to which its parts possess the property. In particular, some scales are **additive** in the sense that the degree to which an object has the property is the sum of the degrees to which the object's proper parts have the property. Clear examples of additive scales are weight and size: the degree to which I am heavy is the sum of (i) the degree to which my right arm is heavy, and (ii) the degree to which the rest of my body is heavy. In contrast, predicates of temperature are not additive but **intermediate**: the degree to which an object is hot is not the sum of the degrees of heat of the object's parts, but rather some value in between these degrees of heat. If my two arms, two legs, torso, and head are all around 98 degrees Fahrenheit, then (thankfully) my body temperature is not around 588 degrees Fahrenheit but around the same 98 degrees. It is easy to imagine scales which behave differently with respect to part-whole relationships. One important alternative possibility is **maximality**. If a scale is maximal, then whenever an object z is fully covered by non-overlapping parts x and y, z's degree on the scale is either the same as x's degree or the same as y's, whichever is greater. Some scales do not seem to encode any such systematic connection, e.g., *happy/sad* and *beautiful/ugly*. How and whether parts and wholes are systematically related constitutes an additional parameter of variation in degree scales.

The issue of part-whole structure is vital for a scalar treatment of deontic modals because many interesting questions in the study of deontic modals are about monotonicity — when and whether embedding under a deontic modal preserves or reverses entailment relations between propositions. In possible worlds semantics entailment is ultimately about part/whole relations, because it is modeled as a subset (part-of) relation between propositions construed as sets of worlds.

Given this, questions about the monotonicity of deontic modals can be refined into questions about the part-whole structure of deontic scales. Both additive and maximal scales have the property that the degree to which an object (proposition) has some property is at least as great as the degree to which its constituent parts (subsets) have the property. Suppose that *ought* means "has moral value greater than θ" for some threshold θ. (Under fairly light assumptions, this subsumes the common "better than not" gloss as a special case.) Then additivity or maximality of deontic scales would entail that *ought* is upward monotonic: if ϕ entails ψ then $ought(\phi)$ entails $ought(\psi)$. This is because the entailing sentence ϕ denotes a subset of the denotation of the entailed sentence ψ,

and maximal and additive scales require a set to have a degree of the relevant property at least as great as any of its subsets.

If deontic scales are either additive or maximal, then, we expect inferences like the following to be valid on the threshold semantics for *ought* just mentioned.

(5) a. You ought to mail this letter.
b. So, you ought to mail this letter or burn it.

Ross' puzzle (generalized from a related issue around imperatives, Ross 1944) is precisely that these inferences are intuitively invalid. Now perhaps this intuitive invalidity can be explained away on Gricean lines as due to the presence of disjunction; if so, an upward monotonic semantics for *ought* remains plausible (Hare 1967; Wedgwood 2006; von Fintel 2012, but see Cariani 2013 for objections). However, Jackson & Pargetter's (1986) Professor Procrastinate scenario provides a structurally identical failure of monotonicity which cannot be explained away by Gricean considerations about disjunction. Here is a version which I think brings out the relevant intuitions more clearly than the original (from Lassiter 2014c, inspired by a scenario in Cariani 2013).

> **Juliet** is considering whether to feign death by taking the drug that Friar Laurence has offered her. If she does, it will put her in a coma, and she will die unless Friar Laurence administers the antidote exactly 10 hours later. If she takes it and the Friar does administer the antidote, she will succeed in convincing her family of her death and she will be able to live happily ever after with Romeo. If she does not take the drug, she will live a long life without Romeo and will be less happy; this is much better than being dead, though. Unfortunately, the Friar is known for being cruel and capricious, and it is extremely likely (though not totally certain) that he will "forget" to administer the antidote if she takes the drug.

Many people find both of the judgments in (6) reasonable in this scenario.

(6) a. It ought to be that Juliet does not take the drug.
b. It ought to be that Juliet takes the drug and the Friar administers the antidote.

But if *ought* is upward monotonic this pattern of judgments should be incoherent. The sentence embedded under *ought* in (6b) entails the negation of the embedded sentence in (6a), and so the truth of (6b) should entail the truth of *It ought to be that Juliet takes the drug* and the falsity of (6a).[2]

This is one of several arguments that have been given in favor of the claim that *ought* is not upward monotonic; see Lassiter to appear: §5 for several more. As a corollary, if we were correct to suppose that the meaning of *ought* is built around a scale of moral/practical value, this scale cannot be either additive or maximal. (By the way, the theories of Lewis (1973) and Kratzer (1981, 1991) make use of qualitative scales that are effectively maximal.)

2 That is, on the usual assumption that *ought*(ϕ) and *ought*($\neg\phi$) cannot both be true. Note that this version of the puzzle problematizes von Fintel's (2012) attempt to explain away Procrastinate cases within a quantificational semantics by supposing that the domain of quantification shifts between the two sentences. The fact that it is specifically mentioned that the Friar *might* administer the drug makes it look pretty *ad hoc* to stipulate that worlds where he does so are not considered in forming this *ought*-judgment.

A third option, if the value scale is neither additive nor maximal, is that it is intermediate. If so, we predict that (6) is a coherent pattern of judgments: $\phi \wedge \psi$ can have value greater than θ while ϕ does not, as long as the value of $\phi \wedge \neg\psi$ (Juliet takes the drug and the Friar does not administer the antidote) is lower than θ. Then, the value of ϕ — equivalently, of the disjunction $(\phi \wedge \psi) \vee (\phi \wedge \neg\psi)$ — will be intermediate between that of $\phi \wedge \psi$ and of $\phi \wedge \neg\psi$, and may well fall below the threshold. For similar reasons, if deontic scales are intermediate we also do not predict the validity of the Ross inference in (5).

But it would be nice to go beyond simply rendering intuitively true judgments like (6) *coherent*, and rendering intuitively false sentences like (5b) *not entailed by obviously true ones*. As Charlow (2013) emphasizes, it would be much better if we could also explain why the former are felt to be *true*, and the latter *false*, in the relevant scenarios. To do this we need to propose a specific intermediate scale for the relevant expressions whose structural properties are related to independently motivated features of the context. We also need to make some assumptions about how the lexical meanings of the various deontic expressions make use of this scale. We can make progress on this front if we adopt the proposal that deontic scales have the structure of **expected value** — including, as a special case for moral judgments, **expected moral value**.[3]

First, we assign a value $V(w)$ to each possible world w in the set of all worlds W (or perhaps to coarser-grained objects such as cells in a partition of the set of possible worlds; the difference will not be crucial here). The value of a world is a real number representing how morally or practically desirable it would be if all of the facts of the world were arranged as in w. The use to which we put value assignments will ensure that they are unique up to positive affine transformation. That is, the numerical value $V(w)$ assigned to a world w is not meaningful per se, but it gains meaning by virtue of the relative sizes of the gaps between $V(w)$ and the values of other worlds $V(w')$ and $V(w'')$.

Second, we assume that a probability measure P is provided ("by context", however this is spelled out). This function maps propositions $A \subseteq W$ to real numbers in the $[0,1]$ interval. This function obeys the usual constraints: it is additive for disjoint propositions, $P(W) = 1$, and its domain is a (σ-)algebra, closed under (countable) union and complement. This could be a probability measure representing the state of knowledge of the speaker or of the holder of an obligation, but we probably need to allow that it could belong to another relevant individual or be abstracted from the information available to a group. It could even be the trivial probability measure of an omniscient being, assigning probability 1 to the unit set containing the actual world and 0 to all others. (This conceit can plausibly be used to model the "objective" use of *ought*: see Wedgwood ms.)

Next we define the *expected value* of a proposition, i.e., the average value of the worlds in the proposition, where each world's value is weighted by the probability that *that world* will be actual on the condition that the proposition is true.

$$\mathbb{E}_V(\phi) = \sum_{w \in \phi} V(w) \times P(\{w\}|\phi)$$

We can derive from this definition a formula for calculating the expected value of a disjunction of disjoint propositions: it is a weighted sum of the expected values of the individual disjuncts,

3 To simplify the math, I'm pretending that the set of possible worlds W is finite. I will also continue to be sloppy about the distinction between sentences and the propositions that they denote relative to a context. It should always be clear in what follows which is intended.

where the weights are given by the probabilities of the disjuncts conditional on the disjunction itself (Jeffrey 1965b: §5).

$$\mathbb{E}_V(\psi \vee \chi) = \mathbb{E}_V(\psi) \times P(\psi|\psi \vee \chi) + \mathbb{E}_V(\chi) \times P(\chi|\psi \vee \chi)$$

Expected value is an intermediate scale. Suppose that ϕ and ψ are mutually exclusive, $\mathbb{E}_V(\phi) > \mathbb{E}_V(\psi)$, and that neither has probability zero conditional on the disjunction $\phi \vee \psi$. Then it is easy to show, as a consequence of the previous equation, that $\phi \vee \psi$ has an expected value intermediate between those of ϕ and ψ.

$$\mathbb{E}_V(\phi) > \mathbb{E}_V(\phi \vee \psi) > \mathbb{E}_V(\psi)$$

If we are seeking an intermediate scale, expected value is a candidate. Why would we think that it is the right choice? One reason is that this semantics effortlessly encodes the empirical phenomenon of **information-sensitivity**: our judgments about the truth or appropriateness of deontic sentences frequently depend on what information is available to the deliberating agent or to us as evaluators of the sentences. (Note however that the expected-value semantics is **not** "seriously information-sensitive" in the sense of Kolodny & MacFarlane (2010): information can reverse preferences over propositions or actions, but not preferences over worlds. As Charlow (2013) discusses, this is a desirable limitation.) Information-sensitivity is an issue that has exercised philosophers considerably in recent years, leading some to weaken their semantic theories considerably or adopt other extraordinary measures. But it is an automatic consequence of the semantics just sketched, for exactly the same reason that information plays a crucial role in Bayesian decision theory. When deciding whether some proposition ought to hold (including the proposition that some agent takes some action), we must consider among other factors how good or desirable the proposition is. In doing so, we should not only take into account the best outcomes that might result if the proposition is true (as the Lewis/Kratzer semantics does), or the worst (as the risk-averse theory of Cariani, Kaufmann & Kaufmann (2013) does). Rather, you must consider all possible outcomes and weight their contribution to the overall result by the probability they they will be the actual outcome if the proposition does occur. This bit of common sense is faithfully encoded by Bayesian reasoning, and we simply port its structure over from theories of subjective decision-making to theories of moral and practical language.

The **Juliet** scenario given above can also be used to illustrate the information-sensitivity of deontic scales. Crucially, the story was arranged so that the expected value of *Juliet takes the pill and the Friar administers the antidote* is very high, and the the expected value of *Juliet takes the pill and the Friar does not administer the antidote* is very low. *Juliet takes the pill* is logically equivalent to the disjunction of these sentences, and so its expected value is a probability-weighted average of the expected value of these sentences. Above we supposed that it was highly likely that the Friar will fail to administer the antidote if Juliet takes the pill, and a reasonable judgment was that Juliet ought not to take the pill. But if we change the story so that the Friar almost certainly *will* administer the antidote, intuitions may shift toward the judgment that Juliet *ought* to take the pill, especially if it is clear from the story that being with Romeo is a great good.

More generally, as we manipulate the continuous range of probabilities for the Friar's actions systematically, intuitions about the relative goodness of Juliet's possible actions seem to vary continuously (cf. Carr 2012; Cariani 2015; Lassiter 2014c). In addition, there is a clear interaction

with the relative goodness of the possible outcomes: the importance of being with Romeo affects the probability needed for a clear judgment that Juliet ought to take the pill.

This account immediately generalizes to the core of the much-discussed **Miners' Puzzle** (Regan 1980; Kolodny & MacFarlane 2010). I will not pause to repeat the prose version of this now-familiar scenario. For reference, Table 1 summarizes the space of actions and outcomes specified in the story.

	Block A	Block B	Block neither
Miners are in A	10 survive	0 survive	9 survive
Miners are in B	0 survive	10 survive	9 survive

Table 1 Outcomes of possible actions, by world-type, in the standard Miners' Puzzle.

As Charlow (2013); Lassiter (2011); Cariani et al. (2013) discuss, the most basic semantic puzzle in the Miners' scenario is not the interaction with conditionals on which Kolodny & MacFarlane focus. A more fundamental problem is how the judgment that *We ought to block neither* is true can be so robust, given that it is clear that the *best* outcomes are all of the following form: either we block A (and the miners are in fact in A) and we block B (and the miners are in fact in B). In either case, all miners survive, while doing nothing ensures that someone will die.[4] On the usual assumption that "ought" means "true in all of the best accessible worlds", this set-up entails that *We ought to do nothing* should be clearly false, and *We ought to either block A or block B* should be clearly true. In the Miners' Puzzle, the thing that we ought to do is guaranteed suboptimal: it is something that we do in *none* of the best worlds.

When we focus on this issue, the puzzle is a variant of Jackson's (1991) Medicine scenario involving a doctor who much choose whether to prescribe a promising but very risky experimental treatment or a safe but mediocre treatment. Even if the best possible result of the experimental treatment is clearly better than the best possible result of the safe treatment, most people have the intuition that the doctor ought to prescribe the safe treatment, given that there is a significant chance that the experimental treatment will (say) cause the patient to die.

4 This analysis is challenged by von Fintel (2012), who claims that the worlds where we do nothing and some miner(s) die actually are better than the worlds where we pick right and they all live. As stated, this judgment strikes me as incredible. But the discussion around it suggests that the judgment is based not on consideration of possible worlds *qua* fully specified states of affairs with no remnant uncertainty, but on sets of worlds where we perform certain actions, where uncertainty about which world is actual remains a relevant consideration. (Kolodny & MacFarlane (2010: §4.3) seem to slip similarly between intuitions about worlds and propositions: "a world in which both shafts are left open may be more ideal than one in which shaft A is closed relative to a less informed state, but less ideal relative to a more informed state".) This is not the question, though: we already knew what ordering on worlds a classical semantics would have to deliver in order to generate the right predictions, but the problem was that it is implausible that a ranking of worlds in terms of (moral) value could give us that ordering. What we need is a rule that can *derive* the relevant judgments about which action is best (taking into account uncertainty about consequences) from a plausible ranking of fully specified worlds. The Bayesian semantics delivers exactly this. However, if we want to complicate the account by taking into account whether individuals with certain kinds of information take certain actions, this is no barrier: see section 4.1 for discussion.

In the Medicine puzzle, as Jackson (1991: 463) points out, "[t]he obvious answer is to take a leaf out of decision theory's book". First, weigh the values of the possible outcomes of the available actions against the probability that they will be actual if the action is taken. Whichever action has the greatest expected value is the one which we ought to take. This procedure is encoded in the Bayesian semantics we have been discussed, and it combines reasoning about values with reasoning about probabilistic information in a common-sensical way. The information-sensitivity of *ought* is predicted immediately, with no need to add additional mechanisms or assumptions to the theory.

The same treatment extends immediately to the Miners' puzzle, and to variants which involve detailed manipulations of probabilities and outcomes which are problematic for non-Bayesian theories (Lassiter 2014c). A plausible analysis of Kolodny & MacFarlane's (2010) *ought*-conditional interactions is also available in terms of conditional expected value. See Lassiter to appear for details.

2 "Controversial normative assumptions" and binding

Carr (2012) and Charlow (2015) object to this project in various ways. Carr argues that all of the usual theories of deontic modals "build controversial normative assumptions into the semantics", and offers a generalized theory that places most of the contentful semantics of *ought* into the contextual parameters. Charlow takes Carr's objection a step further, arguing that we can convert Carr's arguments together with some problems around disagreement into a general argument against a truth-conditional semantics for deontic modals and in favor of a deontic expressivism. Both seem to agree that the Bayesian theory, along with its competitors, confuses the meaning of deontic modals with the way that people reason about obligations.

This criticism is far-reaching if correct. The Bayesian semantics, obviously, builds the formal structure of Bayesian decision theory into the semantics. The semantics advocated by (e.g) Lewis (1973); Kratzer (1991); von Fintel (2012), since it focuses on what happens in the *best* accessible worlds where an action is performed, emulates a decision theory with a MaxiMax choice rule. Cariani et al.'s (2013) alternative effectively enforces a MaxiMin choice rule, paying attention only to what happens in the *worst* accessible worlds where an action is performed.[5]

Carr's empirical focus is information-sensitivity, which she elaborates and uses to argue convincingly that neither the MaxiMax nor the MaxiMin semantics is able to account for a fuller range of variants on the Miners' Puzzle (varying information and outcomes, along the lines discussed

5 Note — as a number of authors have recently pointed out, including Cariani (this volume) and Charlow (this volume) — that these semantic theories do not technically force deontic judgments to emulate (resp.) MaxiMax or MaxiMin decision procedures. But this is for a boring reason: until recently, everybody assumed that deontic judgments were related to something that was independently motivated, such as moral judgments about the relative values of certain fully-specified outcomes, for which uncertainty is not a relevant consideration. If we weaken our metasemantic theory so that the parameter controlling values is not tied to anything that can be independently motivated and examined empirically, we grant ourselves the freedom to reverse-engineer whatever value parameter (ordering source, etc.) is needed to support whatever the observed judgments are in a given context. There does not seem to be any technical barrier to making this move, but if it is the only way to save the formal structure of a best- or worst-worlds semantics, the choice seems clear: a theory which delivers the right judgments on the basis of information about the values of fully-specified outcomes — without requiring extra degrees of freedom — is greatly preferable on general methodological grounds of restrictiveness and predictive power.

briefly above). An obvious response would be to search for a better semantics, as we did in the last section. Carr acknowledges obliquely the possibility of a more empirical successful semantics with a Bayesian character, but does not offer any empirical objections to this project; instead, she moves immediately from the failure of MaxiMax and MaxiMin to the proposal that we should parametrize the semantics by decision rules, allowing that any rational decision theory could be supplied by "context" and used to calculate truth-conditions.[6] In this way the account effectively subsumes the Bayesian semantics as a special case, invoked whenever the context supplies a decision rule based on expected value. Perhaps the context always happens to be like this, in which case the data would *appear* to support the Bayesian theory; but Carr's position entails that this would be merely apparent.

It is striking that Carr moves without argument from the premise that neither of two salient alternative semantic proposals explains information-sensitivity to the conclusion that *no* proposal can solve the problem, unless we greatly increase the expressiveness of the semantics — equivalently, unless we temper the predictive ambitions of our semantics. What is missing is a generic argument showing that no semantic theory of *ought* that encodes substantive normative assumptions could be successful. Without such an argument, we would be justified in insisting on the most empirically restrictive semantics available that is compatible with the available data (which is, I have argued, the Bayesian semantics).

A Moorean Open Question argument might do the trick here: if the meaning of deontic modals encoded controversial normative assumptions, they would not be controversial, since it would be apparent to all that these assumptions are incorrect. Like many before me, I don't find this kind of argument very convincing: it assumes that we are able to reliably introspect meaning facts, and it relies on a sharp distinction between meaning facts and facts about the world (cf. §4.2 below). However, there is a hint of an empirical argument in Carr's (2012: §5) discussion of binding in conditionals, which is elaborated by Charlow (2015: §5.2). Carr points out that it makes sense in the Miners' scenario to say things like

(7) a. If MaxiMax is right, we ought to either block shaft A or block shaft B.
b. If MaxiMin is right, we ought to do nothing.

This can be explained if the meaning of *ought* is sensitive to a decision rule parameter, and the antecedent of this conditional binds this parameter temporarily for the purpose of evaluating the consequent. What's more, it could be taken to show that the meaning of *ought* could not build in a specific decision rule: if it did, conditionals of this form would all be trivial. That is, a conditional whose antecedent describes the decision rule given by the correct semantics would be equivalent to (8a), and any antecedent which describes a decision rule incompatible with the correct semantics would be equivalent to (8b).

(8) a. If $3 = 3$, we ought to perform action X.

6 One could reasonably disagree with the implication that MaxiMax and MaxiMin are examples of rational decision rules. The general point remains that there *might* be multiple rational decision rules which make different recommendations in some situations. Certainly it would be interesting and relevant for empirical semantics if we could find evidence that people vary in the procedures that they use to integrate uncertainty and values in reasoning about obligation, regardless of whether such people would count as "rational".

b. If 3 = 7, we ought to perform action X.

Since this is obviously not a correct prediction, the meaning of *ought* cannot build in any specific decision rule, and the parametrization solution is preferable.

This argument is initially quite plausible, but it is an instance of a very general problem involving conditionals and meaning variation, rather than a special problem for certain accounts of deontic modals. For example, in *Nix v. Hedden* (1893) the U.S. Supreme Court considered the question of whether tomatoes are fruit. The Tariff Act of 1883 placed an import duty on fruits, but not on vegetables, and a family of importers named Nix had sued Hedden, the collector of the Port of New York, arguing that they should not have been forced to pay the import duty on their tomatoes. Botanical usage favored the classification of tomatoes as fruits, while ordinary usage favored classification as a vegetable.

Crucially, both sides of this dispute could happily agree on the following two conditionals: the dispute was about determining which antecedent was true.

(9) a. If tomatoes are fruit, the duty applies to imported tomatoes.
b. If tomatoes are not fruit, the duty does not apply to imported tomatoes.

The same semantic puzzle illustrated for (7) and (8) arises here. I take it that, if tomatoes are fruit, then — for reasons familiar from Kripke (1980) and Putnam (1975) — they are fruit in all (metaphysically) possible worlds. So, we expect the conditionals in (9) to be equivalent **either** to those in (10) **or** to those in (11), *depending on whether or not tomatoes are in fact fruit*.

(10) a. If 3 = 3, the duty applies to imported tomatoes.
b. If 3 = 7, the duty does not apply to imported tomatoes.

(11) a. If 3 = 7, the duty applies to imported tomatoes.
b. If 3 = 3, the duty does not apply to imported tomatoes.

Perhaps this is correct, but if so it is not much help in deciding the case. (In fact, the Court ruled in favor of the Nix family, on the grounds that "the common language of the people" was more relevant than botanical usage to the interpretation of the Tariff Act. Arguments over whether tomatoes are "really" a fruit continue to this day.)

The fact that this disagreement involves a natural kind term is not essential. For example, Ludlow (2014: 78) discusses disagreement about whether the horse Secretariat should have been included in Sports Illustrated's list of "50 greatest athletes of the 20th century". The issue was not whether Secretariat was fast or successful enough, but about whether *athlete* can include non-humans in its extension. Here again, both sides of the debate could agree on the conditionals:

(12) a. If horses can be athletes, then Secretariat belongs on the list of 50 greatest athletes of the 20th century.
b. If horses cannot be athletes, then Secretariat does not belong on the list of 50 greatest athletes of the 20th century.

Plunkett & Sundell (2013) come to a similar conclusion, and point out that a similar argument could be run involving *midwest*, *war*, and many more other ordinary terms.

Standard theories tell us that the conditionals in (9) and (12) must receive a trivializing interpretation, similar to those in (10) and (11). This is not a problem for anyone's theory of the meaning of *ought*, *tomato*, or *athlete*: it is a problem for standard theories of conditionals. What is really going on here is the antecedents of the troublesome conditionals somehow bind the interpretation of certain terms for the purpose of evaluating the consequent. This process can also have side effects by influencing the interpretation of other terms that are semantically related to the term in question. (13) and (14) illustrate.

(13) If Bill ran to the store, Mary positively sprinted. ('We both saw how fast Bill and Mary were moving; if we agree to call Bill's style of movement 'running', then — given the semantic relationship between these terms — we're committed to describing Mary's faster movement as 'sprinting'.)

(14) If Sam is small, Al is tiny. ('We both know how big they are; if we decide to call Sam 'small', then — given the semantic relationship between these terms, and the fact that Al is substantially smaller than Sam — we're committed to calling Al 'tiny'.)

Neither of these conditionals commits the speaker to endorsing the consequent; there may even be an implicature that the speaker thinks the consequent is *not* a good description of the relevant situation.

I predict that a solution to the general problem exemplified by (9), (12), and especially (13) and (14) will account for the binding relationship in (7) automatically. The supposition that the antecedent *MaxiMax is right* is true has an effect on the interpretation of the consequent, via the supposition that a certain semantic issue has been resolved in a particular way. This is parallel to (13) and (14), which can be used even if the speaker is strongly inclined to judge the relevant consequents false. For similar reasons, someone who is implicitly committed to the falsity of the antecedents of (13) and (14)—because they describe an incorrect theory of the meaning of *ought*—is still able to consider the truth-values of *ought*-sentences on the supposition that *ought* is interpreted differently.

How to make sense of the metalinguistic use of conditionals within a compositional model-theoretic semantics is a fascinating question about which I have little to say at the moment. Whatever the best story turns out to be, though, the problem clearly goes beyond the narrow issue of the lexical semantics of deontic modals; trying to resolve the problem in (7) by parametrizing the semantics of *ought* (or *athlete*, *tomato*, or *run*) is missing the generality of the issue. My response here is one which will recur throughout the rest of this essay: critics have mistaken very general semantic puzzles for special problems of particular theories.

3 Causal and evidential expected value

The version of the binding argument that Charlow (2015) gives is interestingly different, because it brings in a deontic analogue of the debate between evidential and causal decision theorists. Modifying a classic example, suppose that drinking milk does not cause obesity, but that there is an undetectable gene *G* which makes teenagers inclined to drink milk and also, in adulthood, makes them obese. Figure 1 depicts this common-cause structure.

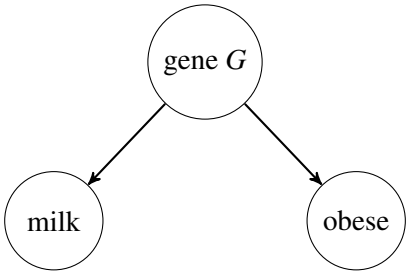

Figure 1 A standard causal model for the milk/obseity example.

If you are a teenager, you don't know if you have G, but you have to decide whether to drink milk or not. Milk is tasty and drinking it will give you great pleasure, whether or not you later become obese; and of course you would strongly prefer not to become obese. Now consider the following:

(15) a. You ought not to drink milk.
b. You ought to drink milk.

If you think that (15a) is true here, I'll say that you have a *evidential-deontic* judgment. This judgment presumably relies on the fact that, if you make the choice to drink milk, this gives *evidence* that you have G, and this is bad news since it means you will become obese.

If you think that (15b) is true, you have a *causal-deontic* judgment about this scenario: since the choice to drink milk has no influence on whether you have G, you should go ahead since it gives you pleasure and does not *causally influence* whether or not you will become obese.

There are several formalizations of causal decision theory in the literature. The rule for calculating expected utility given above was derived from Jeffrey (1965b), and is usually assumed to generate evidential judgments exclusively. A prominent approach to modeling causal judgments, originally proposed by Robert Stalnaker, is to replace Jeffrey's conditional probability-based weighting with a subformula referring to the probability of certain counterfactuals (Gibbard & Harper 1978). Another option, proposed by Lewis (1981), is to weight utilities by certain well-chosen unconditional probabilities. If these formal modifications to the expected-value calculation are the only way to account for the difference between evidential and causal judgments, it seems that at most one of the following conclusions must hold:

a. The correct semantics for deontic modals refers to a scale of expected value calculated using conditional probabilities.

b. The correct semantics for deontic modals refers to a scale of expected value calculated using probabilities of counterfactuals.

c. The correct semantics for deontic modals refers to a scale of expected value calculated using unconditional probabilities.

d. The correct semantics for deontic modals is noncommittal between these options — either by being parametrized by a decision theory (Carr) or because deontic modals can be used to

express, in a non-truth-conditional fashion, the recommendations of these decision theories and many more (Charlow).

Charlow argues that, since native speakers show a mixture of causal-deontic and evidential-deontic judgments, none of (a)-(c) can be right, on pain of rendering these speakers' judgments incorrect *as a matter of meaning*. He concludes that (d) must be the correct approach.

This argument misses the fact that there is another approach to causal decision theory, relying on causal models, which does not require us to modify Jeffrey's formula for calculating expected values. This approach and its implications for the causal/evidential debate is discussed clearly by Meek & Glymour (1994). It is closely related to Pearl's (2000: §4) formulation of causal decision theory, which relies on the addition to the probability calculus of a *do* operator representing an intervention which modifies the probabilistic dependencies among variables in a causal graph. (See Sloman 2005; Sloman & Lagnado 2005 and the papers in Gopnik & Schultz 2007, among many others, for empirical evidence of the relevance of this approach to causal modeling to human cognition.)

In brief, Meek & Glymour (1994) argue that the causal/evidential divide reduces to the choice of how actions are modeled in a given probabilistic model: expected values can be calculated with respect to a measure conditionalized on the *observation* that an action has been performed (evidential) or the existence of an *intervention* designed to produce that action (causal). Expected utilities are calculated in the same way in both cases, but when actions are treated as interventions, the action does not influence the probability of nodes which are not causally dependent on it. In other words, the expected utility of an action *construed as the result of an intervention* will not be influenced by the "news value" of the action about its causes and their other effects: interventions render actions independent of their other causes, and so does not influence their probability or those of other effects which are not also effects of the action in question.

More concretely, in the case of the milk/obseity example: if we model drinking milk as an ordinary event, the choice to drink milk is evidence that you have G, and having G means that you will become obese (Figure 2, left). This means that the Jeffrey expected value of drinking milk will be lowered by the fact that it raises the probability of having G, and so of becoming obese. This gives rise to the evidential-deontic prediction *you ought not to drink milk*.

On the other hand, we could also model milk-drinking as the result of an exogenous intervention, i.e., an uncaused variable $\mathcal{I}_{\text{milk}}$ with three possible values: T, F, and None, where the latter represents no intervention. On Meek & Glymour's (1994) account, setting $\mathcal{I}_{\text{milk}}$ to T or F has two effects: it sets the probability of **milk** to 1 or 0 (resp.), and it renders **milk** independent of G. As a result, the existence of an intervention $\mathcal{I}_{\text{milk}} = T$ is not informative about the probability of the undesirable outcome **obese** (Figure 2, right). (Note that this independence does not follow from the structure of the causal model depicted, but rather from the definition of the full joint distribution: see Meek & Glymour 1994: 1008-9.) In this case, the expected utility of drinking milk is no longer sensitive to the possibility of obesity. Since we are assuming that drinking milk is intrinsically desirable, we wind up with the causal-deontic prediction *you ought to drink milk*.

The difference between an observation and an intervention can be represented in a visually more perspicuous—but equivalent—way by adopting Pearl's (2000) **do** operator, where an intervention triggers a "surgery" which removes edges from the causal graph. This is pictured in Figure 3.

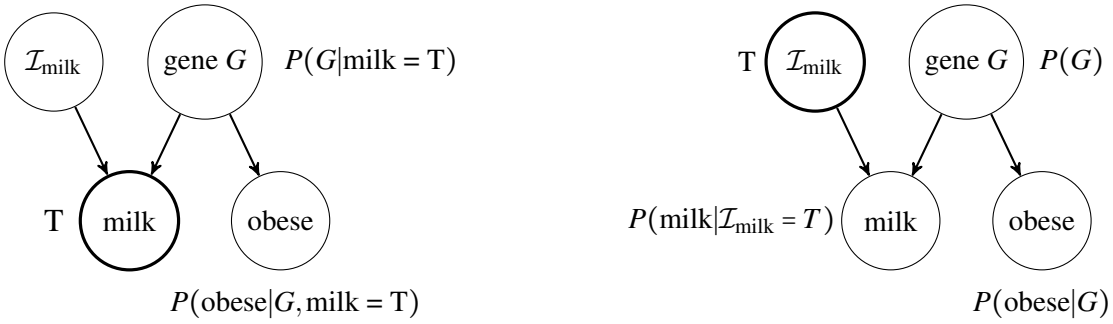

Figure 2 Effect of conditioning on an observation (left) vs. an intervention (right), following Meek & Glymour (1994).

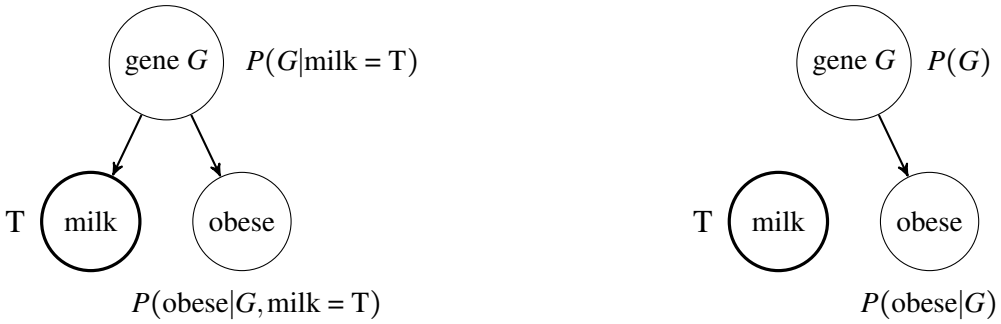

Figure 3 Effect of conditioning on an observation (left) vs. an intervention (right), following Pearl (2000).

Meek & Glymour's (1994) approach to modeling the difference between evidential and causal judgments does not require varying the formula for calculating expected values. Rather, the difference is that we are conditioning on different events: drinking milk, or the existence of an intervention which sets $P(\mathbf{milk}) = 1$. Since the internal structure of the probability distribution is not something that is fixed by the Bayesian theory—like other theories, we treat it as a contextual parameter—the Bayesian theory makes room for both causal-deontic and evidential-deontic judgments. As a result, the observation that there is (or at least may be) variation among speakers in intuitions along these lines is not problematic for the Bayesian theory, and does not support the conclusion that we should weaken the semantic theory in the ways that Carr and Charlow suggest.

For the case of binding specifically, it is natural to analyze Charlow's examples as follows:

(16) a. If causal decision theory is right, you ought to drink milk.
b. If evidential decision theory is right, you ought not to drink milk.

While I am strongly inclined to endorse both the antecedent and the consequent of (16a), I also

think that (16b) is clearly true. Is this a problem? Not at all, if the antecedent *if evidential decision theory is right* is interpreted as "If the right way to model milk-drinking here is to treat it as a caused event, rather than an exogenous intervention". If we require the $P(\cdot)$ parameter to conform to that supposition, (16b) will come out true on the Bayesian semantics. For similar reasons, someone with deontic-evidential judgments should be able to see (16a) as true because he can consider the effects of adopting a probabilistic model in which milk-drinking is treated as an intervention.

4 Variation in judgments and (conceivable) disagreement

Cariani (2015) and Charlow (2015) argue that the Bayesian semantics cannot make sense of certain disagreements involving *ought*. (In Charlow's case, the argument is meant to apply more broadly.) How can we understand disagreement about an *ought*-claim between two individuals who have the same information and values (preferences over fully specified possible worlds)? To the extent that a disagreement is possible, holding values and information fixed, at least one of the parties must not be employing the $\mathbb{E}_V(\cdot)$ function defined in §1 to map probabilities and preferences over worlds to preferences over propositions. Furthermore, it is implausible that such disagreements can only take place when one of the two parties has misunderstood the language; we must allow that a disagreement could occur even when both are correctly employing their linguistic competence in forming the relevant judgment. So, the choice of $\mathbb{E}_V(\cdot)$ as opposed to some other mapping cannot be part of the semantics. The same argument can be run against any contentful semantics for *ought* (in a sense of 'contentful' to be made more precise below). So, no contentful semantics for *ought* can be correct.

Charlow's discussion focuses on causal vs. evidential judgments, but I do not think that these judgments are problematic for the Bayesian theory, for reasons discussed above (§3). In this section I will discuss two further potential empirical problems: the existence of non-consequentialist deontic judgments and disagreements involving them, and the conceivability of further patterns of judgment and disagreement involving deontic modals. §4.1 shows that major classes of empirically motivated non-consequentialist judgments are readily modeled by the Bayesian theory. §4.2 discusses patterns of judgment that *are* ruled out by the Bayesian theory, notably MaxiMax and MaxiMin: perhaps the *conceivability* of these systems is a problem for the Bayesian theory. I will argue that, unless evidence is found that English speakers actually have judgments of this type, the mere conceivability of such judgments is not something that we should build a semantic theory around.

It may be useful to flag a distinction between two kinds of disagreements: reflective theory-laden disagreements among semanticists and/or moral philosophers, and disagreements about specific cases among non-theorists. Some philosophical work on this topic seems to take for granted that both kinds of disagreement are semantically relevant. For example, one of the most frequent criticisms of the Bayesian theory that I have encountered is that it (supposedly) renders the meta-ethical positions of certain philosophers analytically false. Along similar lines, Charlow (this volume) argues at one point that a certain position is not "palatable" because it entails that "your philosophical colleagues" have adopted a position that is trivially true (§5.4). This kind of disagreement is, in my opinion, not very interesting for an empirically-oriented theory of meaning. We are not trying to model the full range of conceivable theoretical positions that someone could take on a topic: the object

of interest is the knowledge that ordinary speakers bring to bear in producing and interpreting language, including grammatical and inferential features of the language as they understand it. I will assume in what follows that the only kind of disagreements that are of interest are those that can or could be motivated by empirical evidence involving non-theorists.

4.1 Non-consequentialist judgments

One kind of disagreement that has been given extensive empirical motivation involves consequentialist and non-consequentialist judgments. A good survey is May (2014), who discusses empirical work on five types of issues along these lines: intentional vs. accidental harm; action vs. omission; means, force, and agents; means vs. side-effect; and personal vs. impersonal harm. While evidence is mixed, non-theorists' judgments frequently aligns with the predictions of non-consequentialist theories. Is this a problem for the Bayesian semantics, as some have suggested?[7]

The answer is "no", and the reasoning is essentially the same as Lewis' in his response to McMichael (Lewis 1978). McMichael had accused Lewis of encoding a problematic utilitarianism into the deontic theory of *Counterfactuals* (Lewis 1973: §5). This objection would be valid if Lewis' theory contained, in addition to a semantic parameter representing deontic value as a world-ordering, some metasemantic restriction on the kinds of information about a world that can go into determining its value. But Lewis alleges no such restriction: in principle, any piece of information about two worlds can play a role in determining their relative goodness. This could include whether Bobby harmed Jimmy intentionally or accidentally, whether he harmed him actively or simply allowed him to be harmed, and so on. These are facts about actions, to be sure, but actions are part of the world, and a semantic theory which has access to information about the whole world will of course have access to information about properties of actions that take place in that world. As Lewis (1978: 85-6) puts it: "The semantic analysis tells us what is true (at a world) under an ordering. It modestly declines to choose the proper ordering. That is work for a moralist, not a semanticist."

Similarly, the Bayesian theory has no difficulty with the fact that people's deontic judgments are frequently sensitive to more fine-grained information about the world than (for instance) the overall balance of good and evil. No metasemantic restriction is in place to prevent the moral value function from ranking worlds in which the harm that Bobby inflicted on Jimmy was intentional below worlds which are identical except that this harm was accidental. The information that distinguishes these worlds involves psychological facts about people in them, but nothing in the theory restricts the moral value function from taking such information into account. If we *wanted* to encode (say) a utilitarian ethical theory into the Bayesian semantics sketched above, we would have to do it by adding a metasemantic restriction which constrains admissible value functions in this way. Even radically nonconsequentalist theories—those in which the consequences of actions are never taken into account—can be modeled within the Bayesian semantics by placing constraints on allowable value functions; all we would have to do is to supply a value function which ignores

7 Note that the claim that it is a problem is not made explicitly by the authors that I am responding to here. However, I have heard it from numerous philosophers in public and private communications on the topic, and I suspect that some assumption along these lines is widespread: judging by his choice of title, even Goble (1996) seems to make it in "Utilitarian deontic logic", one of the first statements of the Bayesian semantics.

the consequences of actions in choosing how to rank two worlds, and pays attention only to the universalizability of agents' actions, their conformance with God's will, or whatever other feature we choose to focus on.

There *are* important limitations to the Bayesian theory's expressiveness, but these limitations involve *logical* relations among sentences. The theory rules out, for example, the possibility that $\phi \vee \psi$ can have value greater than that of ϕ. If God decrees that worlds are classified according to whether people aim to praise Him and to help the poor, He still cannot make praying *or* giving to the poor better than each of the following individually: praying, and giving to the poor. Such logical predictions are the main selling point of the Bayesian semantics. Perhaps there is some strange ethical theory which is not compatible with this prediction; if so, and if we could find non-theorists whose moral judgments coincided with the predictions of this theory, we would be justified in rejecting the Bayesian semantics. As far as I know, no such evidence exists.

4.2 Conceivability arguments and other disagreements

So far we have focused on empirical objections; but there is a well-known philosophical argument which—if successful—would allow us to dispense with empirical evidence and argue from conceivability facts alone. Clearly, a non-theorist could be implicitly committed to Bayesian reasoning about obligation, in the sense that she reasons about obligation and related concepts by considering the full range of possible scenarios, weighing their value against the conditional probability of their occurrence given various actions. (Indeed, I would hypothesize that this is how nearly everyone reasons intuitively about obligation, perhaps excluding people with strong theoretical commitments.) Such a person would be committed to the following judgments about the standard Miners' Puzzle and Jackson's Medicine Puzzle, assuming that information and values are fixed as described above.

(17) a. We ought to block neither shaft.
b. The doctor ought to prescribe the safe but mediocre treatment.

But it is conceivable that there could be a person who is implicitly committed to MaxiMax. Such a person would consider only what could occur in the best possible worlds in which some action is performed, believing that the action with the greatest maximum is what ought to be done, regardless of how likely or unlikely it is that this maximal outcome will be actual if the action is performed. Again assuming that information and values are fixed as above, a MaxiMax enthusiast would make the following judgments:

(18) a. We ought to either block A or block B.
b. The doctor ought to prescribe the risky experimental treatment.

To see how this is a problem, let's define "contentful" more precisely. Along with most of the recent semantic literature on the topic, we have been assuming that the interpretation of *ought* in context is sensitive to an information state and a "value" parameter, either a preference order over worlds $\succcurlyeq^w$ or a value function V. (The latter is a strictly stronger assumption: a value function naturally determines an agreeing world-ordering via the rule $w_1 \succcurlyeq^w s_2 \Leftrightarrow V(w_1) \geq V(w_2)$.) Since it's common ground between me, Charlow, Carr, and Cariani, let's assume that information states include probabilistic information in the form of a measure $P(\cdot)$. We also assume that agents

somehow come to form preferences over objects of a higher type — propositions, or perhaps actions — and that this is what is utilized directly in judgments about (e.g.) *ought*. Depending on the theory, these preferences are represented either by a preference order $\succcurlyeq^s$ over propositions, or by an expected value function $\mathbb{E}_V(\cdot)$.[8] (Again, the latter is stronger: $\mathbb{E}_V(\cdot)$ naturally determines an agreeing preference order via $\phi_1 \succcurlyeq^s \phi_2 \Leftrightarrow \mathbb{E}_V(\phi_1) \geq \mathbb{E}_V(\phi_2)$.) For the specific case of *ought* we can be fairly noncommittal about how this maps into judgments, except to require that a cooperative speaker in ideal conditions will endorse *ought*(ϕ) only if she judges that ϕ is deontically better than all relevant alternatives. "Better" is spelled out in the obvious way using $\succcurlyeq^s$ or $\mathbb{E}_V(\cdot)$, depending on the theory. Now we can define what it means for a semantics for deontic modals to be "contentful":

> A **contentful semantics for deontic modals** is a semantics which places constraints on the kinds of probability measures $P(\cdot)$ and preferences over worlds ($\succcurlyeq^w$ or V) which can be associated with which preferences over propositions ($\succcurlyeq^s$ or $\mathbb{E}_V(\cdot)$).

The Bayesian theory is as contentful as can be: its parameters $P(\cdot)$ and V together fully determine $\mathbb{E}_V(\cdot)$. So is the Lewis/Kratzer semantics, where $\succcurlyeq^s$ is fully determined by $\succcurlyeq^w$, and $P(\cdot)$ is ignored.

A semantics which is not at all contentful would be one which begins with a $P(\cdot)$ and $\succcurlyeq^w$, or a $P(\cdot)$ and V, and allows them to be associated with any $\succcurlyeq^s$ or $\mathbb{E}_V(\cdot)$ (as appropriate). Such a semantics would allow, for example, that *ought* judgments could be associated with a $\succcurlyeq^s$ ordering generated by a MiniMin rule (you *ought* to perform the action with the worst possible worst-case outcome), or a random shuffling of propositions (RandomChoice), or that the proposition-ordering could be intransitive or cyclic. While we are focusing on the back-and-forth between maximally contentful theories (like the Bayesian one) and very weak theories (like Carr's and Charlow's), there is clearly a lot of logical space between the extremes: it could be that the semantics constrains but does not fully determine the relationship between these components.

I don't know of any direct empirical evidence of variation among non-theorists in which choice rules are used in forming *ought*-judgments, holding information and values fixed. (As discussed above, 'values' must be read here to permit value judgments to include all of the information about a world, including issues of personal vs. impersonal harm, action vs. omission, etc.) A prominent class of potential counter-examples here involves varying attitudes toward risk in *ought*-judgments. Such variation is well-motivated empirically in the case of choice behavior, and it's reasonable to speculate that it exists in deontic judgment as well (though direct empirical evidence involving deontic judgment specifically is needed before we can be confident). However, it is generally possible to model moderate risk-seeking and risk-averse choice behavior in expected utility terms,

8 I've used $\succcurlyeq^w$ where Lewis and Kratzer would use a symbol closer to $\preccurlyeq^w$. I find it more natural to use the former for "at least as good as", and the latter for "at most as good as"/"at least as bad as".

Note that in Kratzer's (1991) theory $\succcurlyeq^w$ is not a contextual parameter, but is determined indirectly by a modal base $f : w \to \mathcal{P}(W)$, an ordering source $g : w \to \mathcal{P}(W)$, and a rule for constructing $\succcurlyeq^w$ from f and g: relative to an evaluation world w_0,

$$w_1 \succcurlyeq^w w_2 \Leftrightarrow w_1, w_2 \in \bigcap f(w_0) \ \wedge \{p \in g(w_0) | w_1 \in p\} \supseteq \{p \in g(w_0) | w_2 \in p\}.$$

Any choice of f and g will determine a reflexive and transitive order $\succcurlyeq^w$. In addition, Lewis (1981) proved that, for any reflexive and transitive order $\succcurlyeq^w$ over a subset of W, Kratzer can choose some f and g which delivers exactly that $\succcurlyeq^w$. The theories are thus equally expressive, and the choice to treat $\succcurlyeq^w$ as a parameter or derive it using Kratzer's more elaborate method makes no difference for our purposes.

unless it is at the probability-insensitive extremes of MaxiMin or MaxiMax (Pratt 1964; Arrow 1965, etc.).

That said, there is considerable disagreement among economists about whether the way that expected utility models capture moderate risk-averse and risk-seeking behavior is plausible.[9] Arguments against the Pratt-Arrow treatment of attitudes toward risk could perhaps be converted into a convincing reason to weaken the Bayesian semantics for *ought* and other deontic items, if they can be given parallel empirical motivatation for deontic judgments specifically. If so, the response should not be to adopt a totally noncommittal semantics; instead, the obvious next theoretical move is to look for inspiration in (for example) Kahneman & Tversky's (1979) Prospect Theory, which makes strong empirical predicitions but does not absorb attitudes toward risk into the value function. Alternatively, a demonstration that some individuals' deontic judgments are totally insensitive to non-categorical changes in probability would do the trick — though I doubt seriously that the latter type of evidence will be found.[10]

Lacking convincing empirical evidence for inter-individual variation in choice rules, we are forced to argue from the *conceivability* of certain kinds of disagreements to substantive conclusions about the actual semantics, in Moorean fashion. This much I am willing to concede: it is conceivable that someone could reason deontically using MaxiMax, MaxiMin, or various other choice rules. Here is my attempt to spell out the assumptions that would be necessary for this observation to constitute a compelling argument against any specific semantic proposal. I do this in rather tedious detail, since there are a number of places where one could object.[11]

1. Sub-argument:

    1a. If we can *conceive* of two competent speakers forming contradictory *ought*-involving judgments as a result of employing different choice rules — holding information and values fixed — then it is *possible* for two competent speakers to form contradictory *ought*-involving judgments as a result of employing different choice rules.

    1b. We can conceive of two competent speakers employing different choice rules in forming *ought* judgments, holding information and values fixed.

    1c. So, it is possible [in the sense relevant to (1a)] for two competent speakers to form contradictory *ought* judgments as a result of employing different choice rules. [From 1a and 1b]

2. Sub-argument:

9 Rabin 2000 is a prominent critique, but see also Chetty & Szeidl 2007 for evidence that expected utility models are not as badly off as Rabin's critique might make it appear.

10 There is interesting work suggesting intra-individual variation in judgments triggered by cognitive load manipulations (Greene, Morelli, Lowenberg, Nystrom & Cohen 2008). I assume here that we are modeling reflective judgments formed under minimal task demands. There is also evidence of framing effects in moral judgments, with logically equivalent scenarios being judged as worse when their negative aspects are highlighted (Sunstein 2005; Kern & Chugh 2009). This is sometimes described as "moral loss aversion", but it's qualitatively different from risk-aversion, and it's usually thought to be a judgmental bias rather than an effect that should be predicted by a decision theory.

11 Throughout the argument, "people", "speakers", etc. are implicitly restricted to non-theorists; otherwise several of the premises would be quite implausible, for reasons discussed above.

2a. If two competent speakers form contradictory *ought* judgments as a result of employing different choice rules, this is a substantive disagreement about an *ought* judgment which is generated by the use of different choice rules.

2b. In the sense of *possible* relevant to (1a), *If p then q* entails *If it is possible that p then it is possible that q*.

2c. So, if it is possible [in the sense relevant to (1a)] for two competent speakers to form contradictory *ought*-involving judgments as a result of employing different choice rules, then it is possible [in the sense relevant to (1a)] for there to be a substantive disagreement about an *ought* judgment which is generated by the use of different choice rules. [From (2a) and (2b)]

3. So, it is possible [in the sense relevant to (1a)] for there to be a substantive disagreement about what ought to happen which is generated by the use of different choice rules. [From (1c) and (2c)]

4. Disagreements are either substantive or verbal. Substantive disagreements require that the parties involved use all of the lexical items with the same meaning when describing the issue about which they are disagreeing. If there is any difference in the meanings expressed by the parties to the disagreement when they describe the issue under contention, the disagreement is merely verbal.

5. So, it is possible to vary the choice rule used in the generation of an *ought* judgment without varying its meaning. [From (3) and (4)]

The argument appears to be valid, though it falls short of establishing that *ought* is not contentful, i.e., that *any* choice rule is compatible with the meaning of *ought*. For that, we would need to assume that, for any definable combination of $\langle P(\cdot), \succcurlyeq^w, \succcurlyeq^s \rangle$ — or $\langle P(\cdot), V(\cdot), \mathbb{E}_V(\cdot) \rangle$, depending on the theory we're working with — we can conceive of a disagreement where one party is employing this combination. This would have to include combinations determined by crazy rules like MiniMin and RandomChoice, as well as combinations where $\succcurlyeq^s$ is intransitive or cyclic. I find it more or less impossible to imagine these things, though perhaps this difficulty can be attributed to other factors (say, to constraints imposed by an intuitive theory of others' psychology).

The real problems with this argument are broader: a number of the premises are debatable at best. Consider premise (1a), for example. The plausibility of this premise depends on what sense of "possible" is relevant. Epistemic possibility is a non-starter—it's easy to conceive of things that are epistemically impossible, like the moon being made of green cheese. Metaphysical possibility is a plausible candidate for the intended sense. If "possible" is resolved in this way, then premise (1a) amounts to the assumption that conceivability entails metaphysical possibility—a hotly disputed point, to put it mildly. For example, Putnam (1975) argues that it is conceivable that water is not H_2O, but — given that water is in fact H_2O — it is not metaphysically possible that it is not. Likewise, if conceivability entails possibility then we can learn that zombies are possible, and that a materialist philosophy of mind is false, all without leaving the armchair. (See Yablo 1993; Chalmers 1996 and the papers in Gendler & Hawthorne 2002. Depending on who you ask, these consequences are either interesting discoveries or reductios of the assumption that conceivability

entails possibility.) Moving away from heady metaphysical debates and back to homely linguistic questions, the broader assumption that meaning facts can be reliably discerned by exploring our intuitions about the conceivability of various kinds of disagreements seems dubious. It assumes that our introspective abilities in this domain far exceed what we seems to hold generally for our linguistic knowledge and most other aspects of our cognitive lives.

A second problem occurs in (1a) and at various other points in the argument, involving the notion of a "competent" speaker. The hedge "competent" occurs constantly in the philosophical literature on disagreement, and for good reason: linguistic variation is a constant presence in our lives, and it is crucial, and very difficult, to factor out its effects in making judgments involving meaning and disagreement. But it is hard to see how the vague notion of a "competent" speaker can be cashed out in enough detail to do the work that is being demanded of it here. Consider again the issue of whether tomatoes are fruit. Many modern English speakers are of the opinion that this is a scientific fact — "Tomatoes are not a vegetable, they are really a fruit". This insistence clearly implies that people who consider tomatoes a vegetable are incompetent, either with respect to *tomatoes*, with respect to *fruit*, or both. Others deny that tomatoes are fruit, in full knowledge of the relevant botanical facts. The latter group would presumably judge the former group as incompetent with respect to the use of one or both of these expressions. (As a quick web search will reveal, both groups are well-represented and vocal.) It is hard to imagine that we could resolve this issue by consulting intuitions: both groups will intuit that the other is mistaken.

To make matters worse, there is a third possibility: perhaps the two groups are speaking subtly different languages, and both are correct relative to their own languages. This is clearly what the Supreme Court thought in its decision in *Nix v. Hedden*: the judgment was carefully worded to indicate that both uses were fully legitimate in their respective realms, and that the only question was which usage should be employed when interpreting a specific law. But it would surely be difficult to convince the two parties in the tomato-classification debate that they are not really disagreeing. In other words, in some cases there probably is simply no fact of the matter about whether a disagreement is substantive or verbal in nature: the vague notion of "speaking the same language" can only bear so much theoretical weight (Chomsky 1986; Lassiter 2008). Even if this is wrong, and there is some hidden fact of the matter in this particular case, it is surely not one that we as theorists can reliably intuit.

The same holds, I suspect, for the case of the imagined *ought*-disagreements between speakers who are implicitly committed to different varieties of decision theory. If we were to encounter such a disagreement in the flesh, it would not be clear how to classify it: perhaps one party is right and the other is wrong, or perhaps the two parties are speaking languages which differ ever so subtly. Intuitions about this case are not clear, and we cannot expect to rely on them in adjudicating the issue. Perhaps, as is plausible in the tomato case, there would not even be a fact of the matter.

The "tomato" and "athlete" examples problematize premises (2a) and (4) of the argument about *ought*-involving disagreements as well. Both of these premises presuppose that there is a clear distinction to be drawn between substantive and verbal disagreements. The assumption that such boundaries exist is independently troubling in light of the general problematization of the analytic/synthetic distinction since Quine 1951. The assumption that a disagreement must be *either* one of meaning *or* one of fact presupposes that a such a distinction exists. Worse, the

assumption that we can reliably intuit which is at play in a given case presupposes that we can *reliably distinguish* meaning facts from facts about the world. But even if we suppose, contra Quine, that the analytic/synthetic distinction is real and theoretically useful, the history of philosophical work which uses or criticizes the concept of analyticity demonstrates clearly that our intuitions about how to classify particular cases are not generally reliable.

In sum, the argument from the conceivability of certain kinds of disagreement is very generic in form, and has several problematic aspects that have nothing particular to do with the semantics of deontic modals. As a result, this argument does not cause me to lose sleep. On the other hand, solid empirical evidence of deontic choice-rule disagreements among non-theorists *would* constitute a strong argument for relaxing the semantics to make room for whatever patterns are actually observed. Alternatively, if there were separable populations of speakers who consistently used *ought* in different ways along this parameter, we might be justified in supposing that their languages differ subtly in the meaning of *ought*. In my estimation, the most fruitful direction for research on deontic modals will be oriented toward empirical issues of this type.

This response still leaves important philosophical questions unanswered. What *is* the relationship between (e.g.) the Bayesian semantics and the notion of analyticity? Isn't the Bayesian semantics implausible because, if correct, it would be an analytic truth that it is correct, and so obvious and trivial? In the back-and-forth between philosophers of language and practicing lexical semanticists, there seems to be a fair bit of confusion on this point at a larger scale. For example, Fodor & Lepore (1998) criticize Cruse's (1986) classic, descriptively-oriented lexical semantics textbook, which spends much time describing relationships among lexical items like entailment, antonymy, and synonymy. In their estimation, this work is hopelessly confused because it confuses the meaning of an expression like *hot* with the inferences that people tend to draw from it (e.g., that something that is hot is not cold). They formulate extensive related criticisms of Pustejovsky's (1995) Generative Lexicon theory, a richly structured theory of lexical semantics intended to account for systematic meaning shifts which occur predictably in certain environments and across entire classes of items. (For example, *book*, *CD*, and *file* can be interpreted both as a physical object in *burn the book/CD/file* and as a container of information in *read the book/file* and *listen to the CD*; the latter, but not the former, is compatible with the object being instantiated as data on an iPad.) Fodor & Lepore argue that, like Cruse, Pustejovsky makes the grave error of encoding worldly facts like "books can contain information" as pieces of information *in the lexicon*. This is confused because it is possible to imagine these facts being otherwise, but information that is in the lexicon could not be otherwise (e.g., Fodor & Lepore 1998: 274). Fodor & Lepore argue that the lexicon should contain nothing at all except statements like these: "cat" denotes CAT, "book" denotes BOOK, and (we can extrapolate) "ought" denotes OUGHT.[12]

Reading Fodor & Lepore 1998 alongside Pustejovsky's (1998) reply makes it clear how deep the misunderstanding goes here: lexical semanticists are simply not playing the game that (at least

12 Interestingly, Cariani (2015) and Charlow (2015) both give *ought* more semantic structure than this, and Cariani endorses a probabilistic semantics for epistemic *likely* along the lines of Yalcin 2010, 2012b; Lassiter 2010, 2011. It's not obvious to me that either of these positions could survive the philosophical weapon of mass destruction conjured up by the disagreement argument: it's possible to *imagine* people disagreeing about pretty much anything that you might think to call "meaning". If we take the disagreement argument to its logical conclusion, I suspect that we will eventually be compelled to Fodor & Lepore's bare-bones concept of lexical semantics for all of these expressions.

some) philosophers think that they are. Pustejovsky points out that Fodor & Lepore wish to throw the theory out on purely *a priori* grounds. They are motivated by the correct observation that "meaning is dirty", but offer no alternative account of the empirical generalizations that Pustejovsky is trying to capture. Meaning is kept pristine, but the overall effect is either to re-categorize these facts as "conceptual" or to give up on the project of explaining them. Pustejovsky's argument has a further striking component: he claims that "the human linguistic capacity is a reflection of our ability to categorize and represent the world in particular ways" (Pustejovsky 1998: 290) — in other words, a theory of language and of the lexicon in particular is intimately connected with a theory of concepts, reasoning, and cognition at large. No notion of analyticity is presupposed, and no analytic truths are supposed to follow from a theory of the lexicon.

This is a conception of meaning that is eminently compatible with Quine's (1951) skepticism about analytic truth. Taking Quine's arguments seriously does not entail that we should stop talking about meanings and facts. (Quine himself did plenty of both.) Rather, it means that we should not expect the distinction between them to bear theoretical weight, especially in edge cases. Apparently analytic truths are simply those which are relatively central in our theories of the world, relatively stable across members of a speech communities, and — for psychological reasons — relatively difficult to imagine being otherwise. For linguists, then, an appropriately Quinean response to the unclarity of analytic truth is to keep doing what we have been doing: we go on theorizing about meaning, keeping in mind that our subjects' meaning-related beliefs may not be sharply distinguished from everything else that they believe about the world.[13] As I understand it, this is how lexical semantics is generally done. On the other hand, this methodology does not presuppose the **non-**existence of analyticity either: it could turn out that some information really is so central that it cannot be doubted and/or could not turn out to be otherwise. It could even turn out, as Fodor & Lepore argue, that the boundary between language and the rest of cognition is drawn so narrowly that all of the interesting work that has been done in lexical semantics is really about concepts and reasoning. For linguists like Pustejovsky (1995, 1998), arguments like this, unaccompanied by testable empirical predictions, are rightly met with a shrug.

My attitude toward the debate around deontic modals and decision theoretic concepts — "How much is in the lexicon, as opposed to being part of a concept of obligation or a theory of moral reasoning?" — is similar. I'm not convinced that there is a meaningful boundary to be drawn between meaning, concepts, and reasoning. But maybe there is; if so, then we cannot expect to locate it by examining intuitions about the conceivability of various scenarios. What is needed is a predictive theory of *ought* that is embedded in a good compositional semantics for the whole language, and responsive to the theory of syntax, pragmatics, and other aspects of cognition. We can then judge the result in terms of the ability of the larger whole to predict the empirical evidence available to us, as well as its overall coherence and simplicity.

It might well turn out that there is a determinate answer to the question of what precisely the **meaning** of *ought* is, in the heavy-duty, analytic-truth-generating sense. If so, perhaps *ought* has

13 Obviously linguists working in the Montagovian tradition would be hard-pressed to follow Quine's verificationism, behaviorism, skepticism about modal notions, etc. What I mean is that this attitude fits in with Quine's general philosophy of science, and in particular his belief that philosophy's job is to aid scientific inquiry, not to dictate methods to it.

a skeletal **meaning** that is normatively noncommittal, as Charlow and Carr argue. If so, much of the work that I do under the rubric of "lexical semantics" is probably better categorized as an investigation into the concept OUGHT, or into English speakers' *ought*-involving reasoning habits, or something else of this sort. This would be an interesting discovery, though not an especially troubling one (except, perhaps, when searching for an institutional affiliation). In any case, we will never know until a testable empirical difference is located.

5 Further objections

Cariani (2015) give two further arguments against a Bayesian semantics: one involving attitude ascriptions and one involving dominance reasoning and zero-probability events. In this section I consider these in turn, arguing again that they are special cases of very general issues rather than specific problems for the Bayesian semantics. Given that these puzzles arise independently, and that their general solutions can be expected to extend to the case at hand, they fail to support Cariani's conclusion that a much more powerful semantic theory of deontic modality is needed.

5.1 Attitudes

Consider again the MaxiMax enthusiast in the disagreement above; let's name him "Carl". If asked what the doctor in the Medicine scenario ought to do, Carl would answer that he ought to prescribe the experimental treatment which brings along with it a very high chance of death. This is because the **best possible** outcome of this treatment is a full recovery, which is better than the best possible outcome of the safe but mediocre treatment. We could move probabilities around as much as we like, making it 99.999% certain that the patient will die under the experimental treatment, but crazy Carl won't budge: unless the experimental treatment leads to death with certainty—with probability 1, and in all epistemically possible worlds—his beliefs about what the doctor ought to do are sensitive only to the most optimistic outcome, the tiny chance that the patient will have a full recovery.

Long-suffering Martha shares Carl's information and values, but has come to the reasonable judgment that the doctor ought to prescribe the safe but mediocre treatment. Having endured Carl's arguments for several pages now, she might turn to you and say in exasperation:

(19) Carl thinks that the doctor ought to prescribe the risky experimental treatment. (But he's wrong.)

Cariani (2015: §2.2) poses an interesting puzzle about how to make sense of Martha's statement. Even allowing that Martha's use of *ought* is well-described by the Bayesian semantics of §1, what are the truth-conditions of (19) in her mouth? (20) is no good, for example.

(20) Carl thinks that, of the options available to the doctor, the action with the highest expected moral value is to prescribe the risky experimental treatment.

If Carl knows what expected value is, he will presumably judge *The action with the highest expected moral value is to prescribe the risky experimental treatment* to be false — after all, he and Martha share all of their values and information. So (20) is false, but Martha's statement in (19) is clearly

true; so (20) is not a good paraphrase of (19). Unfortunately, as Cariani points out, a straightforward, standard operator treatment of *thinks* will end up giving us (20) as the interpretation of (19).

What *does* (19) mean? A prima facie reasonable interpretation might be (21):

(21) Carl thinks that, of the options available to the doctor, the action whose best possible outcome is morally best is the risky experimental treatment.

But it would be a mistake to enrich the semantics by allowing *thinks* to shift some additional parameter, and then conclude that the problem has been solved. This could generate (21), but it isn't enough: Carl and Martha would both endorse the content of the embedded clause in (21), too. The real problem is that Carl and Martha don't agree on how *ought* should be interpreted, and neither (20) nor (21) captures this. The full interpretation of (19) has a normative, metalinguistic component:

(22) Carl thinks that, of the options available to the doctor, the action whose best possible outcome is morally best is the risky experimental treatment, *and that* ought *should be used to describe the action whose best possible outcome is morally best*.

This kind of interpretative flexibility in attitude reports is needed on independent grounds. Consider (yet again) the tomato controversy in *Nix v. Hedden*. Port Controller Hedden (the defendant) could have described the position of the plaintiffs like this:

(23) The Nix family think that tomatoes are not fruit. (But they're wrong: tomatoes are fruit, and I was correct to make them pay the import duty.)

Let's allow that, since Hedden classifies tomatoes as fruit, the interpretation function relevant to interpreting his utterances maps *tomatoes* to a subset of the things that it maps *fruit* to.[14] Still, the intended interpretation of (23) is not "the Nix family think that everything that is a tomato in Nix-English is not a fruit in Nix-English". Rather, the issue is about how "fruit" *should be interpreted*. (23) means roughly (24):

(24) The Nix family think that, for the purpose of interpreting the Tariff Act, objects in the extension of "tomatoes" should be treated as being outside the extension of "fruit". But in fact they are wrong: for the purpose of interpreting the Tariff Act, "fruit" should be construed so that everything in the extension of "tomatoes" falls into its extension as well.

Accounting for this interpretation in the context of a broad compositional semantics for attitude verbs is an interesting challenge that I won't try to pursue here. (See Stalnaker 2003; Shan 2010 for relevant discussion and some directions.) But the lesson, once again, is that the problems attributed to the Bayesian semantics are very general problems indeed. A full theory of how (23) receives the normative, metalinguistic interpretation paraphrased in (24) will — I predict — also explain how (19) is interpreted as (22).

14 If you like the alternative interpretation and are convinced that this is metaphysically impossible, modify the example so that one of the Nixes is the speaker, describing Hedden's position that tomatoes are fruit. The argument then goes through as in the main text.

5.2 Dominance reasoning and zero-probability events

Cariani brings up another objection to the Bayesian semantics, drawn from a recent manuscript by Hájek (ms.) on the foundations of probability. Hájek imagines himself throwing an infinitely thin dart at the $[0,1]$ interval, with a uniform distribution over the possible landing points. It must land somewhere in this interval, but — for any particular value x in this interval — the probability that the dart will land on x is 0. Indeed, the probability the dart's landing point will fall within X is zero for many large, even infinite, sets $X \subset [0,1]$. (For example, X might be the set of all rational numbers in this interval.) Hájek uses this example to bring out several important problems for probability theory, such as the ratio definition of conditional probability and the multiplication definition of independence.

The dart example also brings up an interesting puzzle about Bayesian decision theory. The latter encodes a certain highly intuitive form of dominance reasoning as a theorem, but only when possible-but-zero-probability events are excluded. Imagine that Alan Hájek is about to throw an infinitely thin dart at the $[0,1]$ interval. You have to choose between Option 1 and Option 2. If you choose Option 1, the world will continue in its current lousy condition no matter where the dart lands. If you choose Option 2, there will be eternal world peace if the dart lands on a certain real value r, and nothing will happen otherwise.

	Choose Option 1	Choose Option 2
Dart lands on r	lousy state	world peace
Dart does not land on r	lousy state	lousy state

Table 2 Outcomes of possible choices in the dart-throwing example.

Intuitively, it is clear that you ought to choose Option 2. But suppose that world peace has moral value $1,000,000$ and the current state has moral value -100. The expected value of option 1 is $-100 \times 1 + 1,000,000 \times 0 = -100$. The expected value of Option 2 is, well, $-100 \times 1 + 1,000,000 \times 0 = -100$. The zero-probability event of the dart's landing on r has no effect on the calculation, and so, against intuition, it is morally indifferent which option you choose.

Here again, the objection is not really specific to the Bayesian semantics, but rather is a very general issue for the foundations of Bayesian decision theory, and one which we may expect optimistically to resolve by adopting whatever patch ends up being the right one for the theory in general. But we can also say a bit more about the attempt to use an argument of this form to problematize a proposal about the semantics of English.

One general objection to this style of argument is that infinitely thin darts are extremely remote from our everyday experience. If there were an infinitely thin dart in the room, Hájek wouldn't be able to find it, or to hold onto it long enough to throw it. But I personally find it flatly impossible to imagine such a thing: my mental image is just of a really, really thin dart (which wouldn't be enough to get the puzzle going). These observations may sound silly, but they point to a serious issue: we have no reason to expect that linguistic or moral intuitions that have been forged in our ordinary lives should be a reliable guide to what would or ought to happen in scenarios that are physically

impossible, extremely difficult to conceptualize, or otherwise very remote. Hájek's example, while doubtless an interesting issue for the foundations of decision theory, is not obviously relevant to linguistic semantics if the later is construed as part of the study of human cognition. (Someone who doesn't share the latter assumption might still find the dart puzzle compelling, but then why care about linguistic intuitions at all?)

I personally find the objection from remoteness compelling, but there is a more directly linguistic reason to put the issue to the side: English quite generally seems to enforce a flexible but non-zero granularity in the interpretation of numerical expressions (cf. Lewis 1979; Krifka 2007; Sauerland & Stateva 2007; Bastiaanse 2011). For example, *x is 5 feet 10 inches tall* is an intuitively true statement for many actual individuals x; but surely no human being is *exactly* 5 feet 10 inches tall if 'exactly' is construed to rule out deviations of a millionth of a nanometer. This means that, to a rough approximation, non-mathematical English never talks directly about the event of the dart landing at a real value r. Instead, we talk about events such as the dart falling within a certain range $g > 0$ of r, and — in the example at hand — this is *not* a probability zero event. (Compare Lassiter's (2014a) response to Yalcin's (2007) related objection involving the epistemic adjective *possible*.)

In general, we have to distinguish carefully between the partially formalized, often counter-intuitive semantics of the blend of English and math that we use in (e.g.) measure theory and the homely, flexible, imprecise and vague language that we are modeling when we are trying to give a semantics for English. We can use the former to model the latter, but we cannot assume that they are equally fine-grained. Interestingly, this response seems to be consistent with a comment in Hájek ms. stating the conditions under which the "regularity" assumption that is required for dominance reasoning to go through can be maintained (emphasis in original):

> *On the one hand, we can apparently make the set of contents of an agent's thoughts as big as we like; on the other hand, we restrict the attitudes that she can bear to these contents*—they can achieve only a certain fineness of grain.

I take this to suggest that dominance reasoning can be upheld if the language being modeled (non-mathematical English) is restricted so that it is unable to talk about the troublesome zero-probability sets. It seems plausible that this is the case.

6 Conclusion

The Bayesian semantics does a good job of regimenting grammatical and inferential properties of deontic modals which are troublesome for many theories. In particular, it automatically encodes many of the information-sensitivity facts which have generated so much excitement in the recent literature, and gives a clear semantics for widespread gradability among deontic modals. Since *ought* is the most widely discussed deontic item in the philosophical literature, we have focused on it here as well, but there is a great deal to be said about *may, must, should*, and about *good* and its modified forms *as good as*, *(much) better than*, etc.: see Lassiter to appear.

A number of arguments have been given in recent literature to the effect that the Bayesian semantics is not expressive enough to deal with the interaction of *ought* with conditionals and attitudes, with certain kinds of hypothetical disagreements scenarios, and with situations involving

possible events that have zero probability. In each case, the proposed solution is to weaken so that the predictions of the Bayesian semantics can emerge as one special case among many.

While it might turn out eventually that there is a good theoretical or empirical reason to weaken the semantics, I have argued that none of the arguments surveyed here is compelling. We can explain the problematic data in terms of independent factors—some formal, some philosophical, and some involving known but unsolved linguistic problems.[15] Until a compelling empirical objection is found, the Bayesian semantics continues to be a viable theory of the semantics of deontic modals, and — in my opinion — the best theory available in its balance of empirical coverage and theoretical restrictiveness.

References

Arrow, Kenneth J. 1965. *The theory of risk aversion*. Yrjö Jahnssonin Säätiö.

Bastiaanse, Harald. 2011. The rationality of round interpretation. In U. Sauerland R. Nouwen, R. van Rooij & H.-C. Schmitz (eds.), *Vagueness in communication*, 37–50. Springer.

Bierwisch, Manfred. 1989. The semantics of gradation. In M. Bierwisch & E. Lang (eds.), *Dimensional adjectives: Grammatical structure and conceptual interpretation*, 71–261. Springer-Verlag.

Cariani, Fabrizio. 2009. *The semantics of 'ought' and the unity of modal discourse*: University of California at Berkeley dissertation.

Cariani, Fabrizio. 2013. "Ought" and resolution semantics. *Noûs* 47. 534–558.

Cariani, Fabrizio. 2014. Attitudes, deontics and semantic neutrality. To appear in *Pacific Philosophical Quarterly*.

Cariani, Fabrizio. 2015. Deontic modals and probabilities: One theory to rule them all? To appear in M. Chrisman and N. Charlow (eds.), *Deontic Modals*. Oxford University Press.

Cariani, Fabrizio, Stefan Kaufmann & Magdalena Kaufmann. 2013. Deliberative modality under epistemic uncertainty. *Linguistics and Philosophy* 36. 225–259.

Carr, Jennifer. 2012. Deontic modals without decision theory. In *Proceedings of Sinn und Bedeutung 17*, 167–182.

Chalmers, David. 1996. *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Charlow, Nate. 2013. What we know and what to do. *Synthese* 1–33.

Charlow, Nate. 2015. Decision theory: Yes! Truth conditions: No! To appear in M. Chrisman and N. Charlow (eds.), *Deontic Modals*. Oxford University Press.

Chetty, Raj & Adam Szeidl. 2007. Consumption commitments and risk preferences. *The Quarterly Journal of Economics* 122(2). 831–877.

Chomsky, Noam. 1986. *Knowledge of language: Its nature, origin, and use*. Praeger.

Cruse, D. Alan. 1986. *Lexical semantics*. Cambridge University Press.

von Fintel, K. 2012. The best we can (expect to) get? Challenges to the classic semantics for deontic

15 Note, by the way, that this conclusion should not be construed as an argument against Charlow's expressivism. Expressivism is a plausible thesis about deontic and epistemic modals, and probably other aspects of language as well. I don't think that being an expressivist requires one to reject a Bayesian semantics, though: cf. Yalcin 2012a.

modals. Paper presented at the Central meeting of the American Philosophical Association, February 17. http://mit.edu/fintel/fintel-2012-apa-ought.pdf.

Fodor, Jerry A & Ernie Lepore. 1998. The emptiness of the lexicon: Reflections on James Pustejovsky's *The Generative Lexicon*. *Linguistic Inquiry* 29(2). 269–288.

Gendler, Tamar Szabó & John Hawthorne. 2002. Conceivability and Possibility .

Gibbard, Allan & William L. Harper. 1978. Counterfactuals and two kinds of expected utility. In Harper, Stalnaker & Pearce (eds.), *Ifs: Conditionals, belief, decision, chance, and time*, D. Reidel.

Goble, Lou. 1996. Utilitarian deontic logic. *Philosophical Studies* 82(3). 317–357.

Gopnik, Alison & Laura Schultz (eds.). 2007. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Greene, Joshua D, Sylvia A Morelli, Kelly Lowenberg, Leigh E Nystrom & Jonathan D Cohen. 2008. Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107(3). 1144–1154.

Hájek, Alan. ms. Staying regular. Ms., Australian National University.

Hare, R.M. 1967. Some alleged differences between imperatives and indicatives. *Mind* 76(303). 309.

Hay, Jennifer, Chris Kennedy & Beth Levin. 1999. Scalar structure underlies telicity in 'degree achievements. In *Semantics and Linguistic Theory 9*, 127–144.

Jackson, Frank. 1991. Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics* 101(3). 461–482.

Jackson, Frank & Robert Pargetter. 1986. Oughts, options, and actualism. *The Philosophical Review* 95(2). 233–255.

Jeffrey, Richard C. 1965a. Ethics and the logic of decision. *Journal of Philosophy* 62(19). 528–539.

Jeffrey, Richard C. 1965b. *The logic of decision*. University of Chicago Press.

Kahneman, Daniel & Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* 263–291.

Kamp, Hans. 1975. Two theories about adjectives. In E. Keenan (ed.), *Formal semantics of natural language*, 123–155. Cambridge University Press.

Kennedy, Chris. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30(1). 1–45.

Kennedy, Chris & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2). 345–381.

Kern, Mary C & Dolly Chugh. 2009. Bounded ethicality the perils of loss framing. *Psychological Science* 20(3). 378–384.

Kolodny, Niko & John MacFarlane. 2010. Ifs and oughts. *Journal of Philosophy* 107(3). 115–143.

Krantz, David H., R. Duncan Luce, Patrick Suppes & Amos Tversky. 1971. *Foundations of Measurement*. Academic Press.

Kratzer, Angelika. 1981. The notional category of modality. In Eikmeyer & Rieser (eds.), *Words, Worlds, and Contexts*, 38–74. de Gruyter.

Kratzer, Angelika. 1991. Modality. In Arnim von Stechow & Dieter Wunderlich (eds.), *Semantics: An international handbook of contemporary research*, 639–650. de Gruyter.

Krifka, Manfred. 2007. Approximate interpretation of number words: A case for strategic communication. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 111–126. Koninklijke Nederlandse Akademie van Wetenschapen.

Kripke, Saul. 1980. *Naming and necessity*. Harvard University Press.

Lassiter, Daniel. 2008. Semantic externalism, language variation, and sociolinguistic accommodation. *Mind & Language* 23(5). 607–633.

Lassiter, Daniel. 2010. Gradable epistemic modals, probability, and scale structure. In Nan Li & David Lutz (eds.), *Semantics & Linguistic Theory (SALT) 20*, 197–215. CLC Publications.

Lassiter, Daniel. 2011. Measurement and Modality: The Scalar Basis of Modal Semantics. Ph.D. thesis, New York University.

Lassiter, Daniel. 2014a. Epistemic comparison, models of uncertainty, and the disjunction puzzle. *Journal of Semantics* Advance Access 1–36. doi:10.1093/jos/ffu008.

Lassiter, Daniel. 2014b. Graded modality. To appear in L. Matthewson, C. Meier, H. Rullmann, & E. Zimmermann (eds.), *Companion to Semantics*, Wiley.

Lassiter, Daniel. 2014c. Modality, scale structure, and scalar reasoning. *Pacific Philosophical Quarterly* 95(4). 461–490.

Lassiter, Daniel. 2015. Adjectival modification and gradation. In Shalom Lappin & Chris Fox (eds.), *Handbook of Contemporary Semantic Theory*, Wiley-Blackwell 2nd edn.

Lassiter, Daniel. to appear. *Measurement and Modality: The Scalar Basis of Modal Semantics*. Oxford University Press.

Lewis, David. 1973. *Counterfactuals*. Harvard University Press.

Lewis, David. 1978. Reply to McMichael. *Analysis* 38(2). 85.

Lewis, David. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic* 8(1). 339–359. doi:10.1007/BF00258436.

Lewis, David. 1981. Causal decision theory. *Australasian Journal of Philosophy* 59(1). 5–30.

Ludlow, Peter. 2014. *Living words: Meaning underdetermination and the dynamic lexicon*. Oxford University Press.

May, Joshua. 2014. Moral judgment and deontology: Empirical developments. *Philosophy Compass* 9(11). 745–755.

Meek, Christopher & Clark Glymour. 1994. Conditioning and intervening. *The British journal for the philosophy of science* 45. 1001–1021.

Morzycki, Marcin. 2012. Adjectival extremeness: Degree modification and contextually restricted scales. *Natural Language & Linguistic Theory* 30(2). 567–609.

Morzycki, Marcin. to appear. *Modification*. Cambridge University Press.

Pearl, Judea. 2000. *Causality: Models, reasoning and inference*. Cambridge University Press.

Plunkett, David & Tim Sundell. 2013. Disagreement and the semantics of normative and evaluative terms. *Philosopher's Imprint* 13. 1–37.

Portner, Paul & Aynat Rubinstein. 2014. Extreme and non-extreme deontic modals. To appear in M. Chrisman and N. Charlow (eds.), *Deontic Modals*. Oxford University Press.

Pratt, John W. 1964. Risk aversion in the small and in the large. *Econometrica: Journal of the Econometric Society* 122–136.

Pustejovsky, James. 1995. *The Generative Lexicon*. MIT Press.

Pustejovsky, James. 1998. Generativity and explanation in semantics: A reply to Fodor and Lepore. *Linguistic Inquiry* 29(2). 289–311.

Putnam, Hilary. 1975. The meaning of "meaning". In *Mind, language and reality: Philosophical papers, volume 2*, Cambridge University Press.

Quine, Willard van Orman. 1951. Two dogmas of empiricism. *The Philosophical Review* 60(1).

Rabin, Matthew. 2000. Risk aversion and expected-utility theory: A calibration theorem. *Econometrica* 68(5). 1281–1292.

Regan, Donald. 1980. *Utilitarianism and Co-operation*. Oxford University Press.

Ross, Alf. 1944. Imperatives and logic. *Philosophy of Science* 30–46.

Rotstein, Carmen & Yoad Winter. 2004. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics* 12(3). 259–288.

Sassoon, Galit W. 2013. A typology of multidimensional adjectives. *Journal of Semantics* 30(3). 335–380.

Sauerland, Uli & Penka Stateva. 2007. Scalar vs. epistemic vagueness: Evidence from approximators. In M. Gibson & T. Friedman (eds.), *Proceedings of semantics and Linguistic Theory xvii*, CLC Publications, Cornell University.

Shan, Chung-chieh. 2010. The character of quotation. *Linguistics and Philosophy* 33(5). 417–443.

Sloman, Steven A. 2005. *Causal models: How we think about the world and its alternatives*. Oxford University Press.

Sloman, Steven A & David A Lagnado. 2005. Do we "do"? *Cognitive Science* 29(1). 5–39.

Stalnaker, Robert. 2003. On considering a possible world as actual. In *Aristotelian society supplementary volume*, vol. 75 1, 141–156.

Sunstein, Cass R. 2005. Moral heuristics. *Behavioral and brain sciences* 28(4). 531–541.

Wedgwood, Ralph. 2006. The Meaning of "Ought". *Oxford studies in Metaethics* 1. 127–60.

Wedgwood, Ralph. ms. Objective and subjective *ought*. Manuscript, University of Southern California. http://www-bcf.usc.edu/~wedgwood/Objective_subjective_ought.pdf.

Yablo, Stephen. 1993. Is conceivability a guide to possibility? *Philosophy and Phenomenological Research* 53(1). 1–42.

Yalcin, Seth. 2007. Epistemic modals. *Mind* 116(464). 983–1026.

Yalcin, Seth. 2010. Probability operators. *Philosophy Compass* 5(11). 916–937.

Yalcin, Seth. 2012a. Bayesian expressivism. *Proceedings of the Aristotelian Society* 112. 123–160.

Yalcin, Seth. 2012b. Context probabilism. In M. Aloni, V. Kimmelman, F. Roelofsen, G. W. Sassoon, K. Schulz & M. Westera (eds.), *Logic, language and meaning*, vol. 7218 Lecture Notes in Computer Science, 12–21. Springer.